学校编码：10384　　分类号_____密级_____
学号：15420101151882　　UDC_____

厦门大学

硕 士 学 位 论 文

基于相对误差的分位数回归的统计推断

Statistical Inference for Relative Error-based Quantile Regression

张楠溪

指导教师姓名：朱建平 教授

专 业 名 称：数量经济学

论文提交日期：2013 年 4 月

论文答辩时间：2013 年 5 月

学位授予日期：

答辩委员会主席：

评 阅 人：

二〇一三年四月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

2013 年　月　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

2013 年 4 月　　日

摘　要

线性回归模型是一种十分常见的模型，常用于揭示因变量与一个或多个解释变量之间的线性关系。由于变量间的线性关系比非线性关系更易于进行拟合，并且由前者的回归模型得到的参数统计量的统计性质更易于进行推断，因而线性回归模型在学术界得到了完善而严格的研究，在实际生活中也得到了广泛的应用。

线性回归模型的参数估计方法种类繁多。本文首先简要介绍了这些参数估计方法的提出背景和构造思想衍变。不论是最小化离差平方和，还是最小化绝对离差和，这些方法的构造仅仅涉及到了目标值与估计值的绝对误差。而对于工程问题，或股票交易数据，这时采用相对误差的思想进行参数估计更为合适。但有时仅仅采用一种形式的相对误差进行参数估计准则的构造欠妥，特别是对于目标值与估计值相差很大的情形。因此，基于两种形式的相对误差的估计准则应运而生。

但线性回归模型并不是万能的。由于其本质是一种条件均值模型，因此线性回归模型并不能对数据的分布给出完整描述。随后，本文介绍了分位数回归模型及其传统的参数估计方法，以及相关流行的算法。并指出这种模型可以完整地模型化分布的位置变化和形状变化，也可将其看作是一种参数估计准则，一种对残差绝对值进行不同权重加权求和的参数估计法。此外，文章还展示了分位数回归模型在众多领域中的丰富应用，以及其本身方法论在处理不同特点数据时的发展状况。

其次，文章提出了基于两种形式的相对误差的乘法分位数回归模型的参数估计准则。并在一定的假设条件下，对由这种估计准则求得的估计量的一致性以及渐进正态性进行了理论证明。随后，文章提到了采用随机加权法来规避渐进协方差矩阵中随机干扰项的密度函数的估计。

最后，文章对这种参数估计准则以及传统的分位数回归参数估计准则的估计效果进行了计算机模拟实验对比和研究，并将这两种方法应用于经典的恩格尔家庭预算数据。不论是实验模拟还是实例计算，结果表明，对于乘法分位数回归模型，本文提出的基于两种形式的相对误差参数估计量具有良好而稳定的表现。

关键词: 分位数回归　乘法回归模型　最小相对误差

Abstract

Linear regression models are very common models, which are used to reveal the linear relationship between response variable and one or several explanatory variables. Since the linear relationship between variables is easier to fit than nonlinear one, and the statistical properties of the resulting estimators are easier to determine, linear regression models were studied thoroughly and strictly by researchers, and have extensive practical applications.

There are various parameter estimation methods of linear regression models. First, backgrounds and ideas are introduced in this paper. The method of both least squares and least absolute deviation is established based on the idea of the absolute difference between target value and estimated value. But relative error-based parameter estimation is sometimes proper, such as engineering problems and stock trading data. And estimating parameters using only one type of relative error is inappropriate for the situation when the difference of target value and estimated value is large. So criteria based on two types of relative error occur.

However, linear regression models are not universal since they are conditional mean models which can not give a complete description of distribution. Then quantile regression models are introduced with their parameter estimation methods and relative algorithms. It is pointed out that this kind of model can model both the distribution change and the shape change entirely. And it can be also treated as a kind of estimation criterion, which is based on the sum of different weight of absolute residuals. Besides, various applications and development of methodology of quantile regression models are showed.

Moreover, the parameter estimation criterion based on two types of relative error is proposed for multiplicative quantile regression models. Then the consistency and asymptotic normality of the estimator are proved under some assumptions. And in order not to estimate error density function in asymptotic covariance matrix, a random weighting method is revived.

Finally, the performance of proposed estimator and traditional quantile regression model estimator are compared in both simulation and real example of classical Engel

budget data. The results show that the proposed estimator enjoys a good and stable performance.

Key Words: Quantile Regression; Multiplicative Regression Model; Least Relative Error

目　录

Table of Contents

第一章　绪论

1.1　研究背景

线性回归是一种将响应变量与一个或多个解释变量之间关系进行建模的方法。线性回归模型是一种很常见的回归模型。由于变量间的线性关系比非线性关系更容易进行拟合，并且得到的估计量的统计性质更容易进行推断，因而线性回归模型得到了较完善而严格的研究，并在实际生活中的应用十分广泛，它也因此成为统计学专业接触学习并应用的第一个模型。

线性回归模型的参数估计方法有许多种。我们知道，经典线性回归模型的分析需要满足一系列的条件，但在实际应用中，数据往往不能达到如此苛刻的条件，因此对于不能满足经典假设条件的一些情形，学者们相继提出了一些不同的参数估计方法。这些方法的构造产生于不同的背景，基于不同的思想，都拥有一些优点和缺点。不论是经典的最小二乘法，还是后来提出的工具变量法、岭回归、主成分分析等方法，都得到了比较深入的研究和广泛的应用。

在参数估计方法的发展过程中，值得一提的是，不论是最小化离差平方和，还是最小化绝对离差和，构造这些方法的思想都仅仅涉及到了目标值与估计值的绝对误差。而在处理一些实际数据，特别是一些工程问题，或者是股票交易数据等，当认为损失和相对误差成一定比例时，采用基于相对误差的参数估计方法就显得更为合适。相对误差不仅能清晰反映绝对误差与目标值（或估计值）的差距，而且还可以消除数据的量纲。起初，学者们提出的相对误差仅仅是绝对误差与目标值或者与估计值二者之中的一种比值，但后来发现这种单一的度量方法并不十分合理。因此，有些学者就将这两种类型的相对误差巧妙地结合起来作为一种新的参数估计准则。如选取二者中的较大值，或者对不同权重下的两种相对误差进行求和等。

线性回归模型是一种广泛使用的统计模型，但它也不是万能的。我们知道，线性回归模型的本质是一种条件均值模型，条件均值并不能完整地反映数据的条件分布特征。与此同时，常用的线性回归模型的参数估计方法，如最小二乘法，

对于数据中的离群值十分敏感。因此如果我们草率地将离群值删除后就对剩余的数据进行线性回归模型拟合的话，利用拟合后的模型来解释实际问题就不能得到一个令人满意的结论。

在这种背景下，Koenker 和 Bassett（1978）[1]提出了分位数回归模型。具体而言：令 Y 是一个实值随机变量，其特征可以通过其右连续的分布函数 $F_Y(y)=P(Y\le y)$ 来完全刻画。那么对于任意的 $\tau\in[0,1]$，Y 的 τ 分位数可以表示为 $Q_Y(\tau)=F_Y^{-1}(\tau)=\inf\{y:F_Y(y)\ge\tau\}$。那么，$\tau$ 分位数的线性回归模型可写为 $Y_i=\mathbf{X}_i^T\boldsymbol{\beta}_\tau+\varepsilon_i,\ i=1,...,n.$，其中 $\mathbf{X}_i$ 是 p 维的解释变量，Y_i 是响应变量，$\boldsymbol{\beta}_\tau$ 是包含 τ 分位数的截距项的回归系数，ε_i 是独立于 $\mathbf{X}_i$ 的随机干扰项。

这种模型比中位数回归模型更加一般化，它将分布的位置变化和形状变化模型化，刻画了基于 X 的 Y 的条件分布。对于某 $\tau\in[0,1]$，损失函数定义为 $\rho_\tau(y)=y(\tau-I(y<0))$。则回归系数 $\boldsymbol{\beta}_\tau$ 可通过最小化损失函数 $\sum_{i=1}^{n}\rho_\tau(Y_i-\mathbf{X_i^T}\boldsymbol{\beta}_\tau)$ 得到。

与最小二乘法相比，分位数回归的参数估计并不如前者容易。不同的参数估计准则形式导致了最小二乘法可以转换为一个简单的线性代数问题，而分位数回归的参数估计只能等价地转换为一个线性规划问题，并借助计算机的辅助来进行计算。相关的算法有单纯形法，内点算法，平滑算法等。

从该模型思想的提出以来，分位数回归模型已广泛应用于各个领域，如医学、生态学、经济金融学等。同时在处理不同特点的数据，如在处理删失数据、离散响应数据、面板数据、结构分位数等问题时，分位数回归的方法也得到了不同程度的发展。

1.2 研究目的和创新

本文的研究目的在于针对乘法分位数回归模型的参数估计提出一种基于两种形式相对误差的估计方法。从理论上证明了这种估计方法求得的估计量的一致性和渐进正态性，并从计算机模拟实验中证明了它的可行性。

本文的创新之处在于将分位数回归与基于两种形式相对误差的估计准则相

结合，应用于乘法回归模型的参数估计中来，不仅可以得到分布的完整描述，还能对目标值与估计值之间的差距做出合理的度量。

1.3　文章内容安排

文章其余部分的内容安排如下：

第二章介绍了线性回归模型及其多种参数估计方法，总结了它们的优缺点；随后介绍了分位数、分位数函数，分位数回归模型和其参数估计准则，以及相关算法；最后列举了分位数回归模型在不同领域中的应用，以及在处理不同特点数据下的发展情况。

第三章提出了乘法分位数回归模型的最小相对误差参数估计法，从理论上探讨了在一定假设条件下，这个估计量满足一致性和渐进正态性。

第四章将本文提出的参数估计法和传统的分位数回归模型的参数估计法进行了计算机模拟实验对比，随后将这两种方法应用于经典的恩格尔家庭预算数据，并进行了相关分析。

第五章给出了本文的最后结论，说明了本文提出的估计方法的可行性，并提出了一些未来有待研究的问题。

第二章　理论阐述和文献综述

在本章中，我们将简要介绍线性回归模型、分位数回归模型的基本概念，它们的参数估计方法，相关计算方法。随后，简要展示了分位数回归模型在众多领域中的广泛应用，以及其本身对于处理不同特点数据的发展情况。

2.1　线性回归模型及其参数估计方法

2.1.1　线性回归模型

线性回归是一种对响应变量与一个或多个解释变量之间关系进行建模的方法。由于变量间的线性关系比非线性关系更容易进行拟合，并且得到的估计量的统计性质更容易进行推断，因而线性回归模型得到了较完善而严格的研究，并在实际生活中获得了广泛应用。通常意义上的线性回归指的是响应变量的条件均值是已知解释变量的仿射函数。但也有其他的考量。如响应变量的条件中位数，或者其他分位数等。

线性回归模型假定响应变量Y_i和解释变量X_i之间是线性关系，并通过一个独立于X_i的随机干扰项ε_i进行建模。模型如

$$Y_i = \mathbf{X}_i^T\boldsymbol{\beta} + \varepsilon_i, \quad i = 1,...,n \tag{2.1}$$

其中，$\mathbf{X}_i = (1, X_{2i},..., X_{pi})^T$，$\boldsymbol{\beta}$是$p$维的包含截距项的回归系数，如$\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$。

我们也可将总体回归模型表示为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

那么，总体回归方程为

$$E(\mathbf{Y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

其中，

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{(n\times 1)}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}_{(p\times 1)}, \mathbf{X} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{p1} \\ 1 & X_{22} & X_{32} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{pn} \end{bmatrix}_{(n\times p)}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{(n\times 1)}$$

$$E(\mathbf{Y}\mid\mathbf{X})=\begin{bmatrix} E(Y_1\mid X_{21},X_{31},\cdots,X_{p1}) \\ E(Y_2\mid X_{22},X_{32},\cdots,X_{p2}) \\ \vdots \\ E(Y_n\mid X_{2n},X_{3n},\cdots,X_{pn}) \end{bmatrix}_{(n\times 1)}$$

样本回归模型可以表示为

$$\mathbf{Y}=\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{e}$$

样本回归方程为

$$\hat{\mathbf{Y}}=\mathbf{X}\hat{\boldsymbol{\beta}}$$

其中，

$$\hat{\mathbf{Y}}=\begin{bmatrix}\hat{Y}_1\\ \hat{Y}_2\\ \vdots\\ \hat{Y}_n\end{bmatrix}_{(n\times 1)},\hat{\boldsymbol{\beta}}=\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\\ \vdots\\ \hat{\beta}_p\end{bmatrix}_{(p\times 1)},\mathbf{e}=\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{bmatrix}_{(n\times 1)}$$

线性回归分析需要满足一些前提条件。下面来回顾下经典线性回归模型必须满足的假定条件。如下：

1. 假定随机干扰项$\boldsymbol{\varepsilon}$是零均值的，即

$$E(\boldsymbol{\varepsilon})=\mathbf{0}.$$

2. 假定随机干扰项$\boldsymbol{\varepsilon}$是同方差的，且无序列相关，即

$$Var(\boldsymbol{\varepsilon})=\sigma^2\mathbf{I}_n$$

其中$\mathbf{I}_n$为n阶单位矩阵。

3. 假定随机干扰项$\boldsymbol{\varepsilon}$和解释变量$\mathbf{X}$相互独立，即

$$E(\mathbf{X}^{\mathrm{T}}\boldsymbol{\varepsilon})=\mathbf{0}$$

除此之外，我们还通常假定解释变量$X_i(i=2,...,p)$是非随机变量。

4. 假定解释变量之间无多重共线性，即

$$\mathrm{Rank}(\mathbf{X})=p$$

5. 假定随机干扰项$\boldsymbol{\varepsilon}$满足正态性，即

$$\boldsymbol{\varepsilon}\sim N(\mathbf{0},\sigma^2\mathbf{I_n})$$

这个假定不是线性回归分析的必需假定。

2.1.2 参数估计方法

线性回归模型的参数估计方法有许多种。但在处理实际问题时，通常发现真实数据并不能完全满足上述一系列假定。那么，针对不同特点的数据，学者们提出了相应的参数估计方法。这些方法的构造产生于不同的背景，基于不同的思想，都拥有一些优点和缺点。下图 2.1 展示了其中一些方法的发展历程。[①]

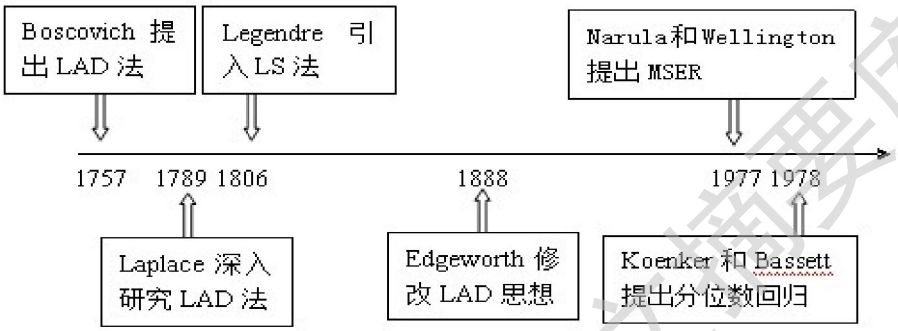

图 2.1 一些参数估计方法的发展历程

最小二乘法（Least Squares，LS）[②]是最简单且最常见的一种估计方法。它的构造理念简洁，而且计算简单。对于模型（式 2.1），LS 的估计准则是最小化离差平方和：$\min_{\hat{\boldsymbol{\beta}}\in\mathbf{B}} \sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\hat{\boldsymbol{\beta}})^2$。求得的估计量形如 $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$。值得一提的是，由这个准则求得的估计量是最佳线性无偏估计量（Best Linear Unbiased Estimator，BLUE）[③]。特别是在满足假定 5 的情况下，即当误差项服从正态分布时，这个估计量达到了模型的 Cramér-Rao 下界，因此是有效的。

广义最小二乘法（Generalized Least Squares，GLS）是 LS 法的拓展。当不满足假定 2 时，即随机干扰项 $\boldsymbol{\varepsilon}$ 存在异方差或者序列相关时，就可以采取这种方法对参数进行估计。其基本思想是对原始数据进行一定的线性变换，使变换后的

[①] 此图参考于 NORTHERN ARIZONA UNIVERSITY 的 Pin Np 的课件“WHAT ARE WE REALLY ESTIMATING IN REGRESSION ANALYSES? An Informal Introduction to Quantile Regression”(P40)，并进行了增添。

[②] 一说是 Adrien Marie Legendre 在其 1806 年发表的书的附录中提出了 LS 法，而 Gauss 声称他在 1795 年就开始使用这个方法。

[③] Gauss-Markov 定理：在线性模型的经典假设下，参数的最小二乘估计量是线性无偏估计量中方差最小的估计量（BLUE）。

数据满足经典假设条件。当随机干扰项$\boldsymbol{\varepsilon}$的方差阵中除对角线外元素均为零时，即随机干扰项$\boldsymbol{\varepsilon}$是异方差但不相关时，可采用 GLS 的一种特殊情形——加权最小二乘法（Weighted Least Squares，WLS）对参数进行估计。同样，GLS 估计量也是 BLUE。

工具变量法（Instrumental Variables，IV）是在处理不满足假定 3，即解释变量与随机干扰项$\boldsymbol{\varepsilon}$相关时的一种参数估计方法。其基本思路是：当解释变量与随机干扰项高度相关时，设法找到另一个与解释变量高度相关，而与随机干扰项不相关的随机变量，用其替代解释变量，从而求得模型参数的估计量。

岭回归（Ridge Regression）估计法是在处理不满足假定 4，即解释变量间存在多重共线性问题时的一种参数估计方法。对于线性回归模型（式 2.1），回归系数$\boldsymbol{\beta}$的岭估计可定义为$\hat{\boldsymbol{\beta}}(k)=(\mathbf{X}^{\mathrm{T}}\mathbf{X}+k\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$，其中$k>0$是岭参数。目前采用较多的选取岭参数的方法是 Hoerl-Kennard 公式法和岭迹法。显然，岭参数取值不同时，得到的估计量是不同的，因此岭估计是一个估计类。并且从其估计公式可知，岭估计是有偏估计。

主成分（Principal Component）估计法也是一种处理解释变量间存在多重共线性的方法。其基本思想是对原始解释变量进行正交变换获得新的解释变量，即主成分，剔除对应特征值比较小的主成分后，对剩余的主成分进行最小二乘回归，最后返回到原来的参数，就得到了主成分分析。主成分估计也是一种有偏估计。

最小绝对值离差法（Least Absolute Deviation，LAD），其基本思想是通过最小化绝对值离差和$\min\limits_{\hat{\boldsymbol{\beta}}\in\mathbf{B}}\sum\limits_{i=1}^{n}|Y_i-\mathbf{X}_i^T\hat{\boldsymbol{\beta}}|$来估计回归系数。求得的 LAD 估计量比 LS 估计量更稳健，尤其不会受到样本离群值的干扰，这使得对于处理那些不能随意删失离群值的数据，LAD 是一种更合适的估计方法。但对于已知的数据集，LS 可求得唯一解，而 LAD 方法求得的解可能并不唯一。虽然 LAD 方法的构造思路简洁，但最小绝对离差线并不容易快速计算。由于 LAD 回归没有类似 LS 的解析解，所以常用一些迭代算法来进行计算。

需要指出的是，分位数回归（Quantile Regression，QR）也可看作是一种参数估计方法。QR 估计出的参数是基于已知解释变量下响应变量的条件分位数，